Spatial Prediction Using Combined Sources of Data

Nancy J. McMillan,[1] David M. Holland,[2] Michele Morara,[1] and Gregory S. Young[1]

(1) Battelle, Measurement and Data Analysis Sciences, Columbus, OH
(2) U.S. Environmental Protection Agency, Office of Research and Development, National Exposure Research Laboratory, Research Triangle Park, NC

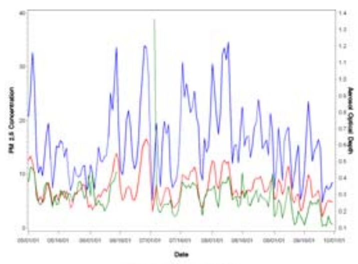
Objectives and Methods

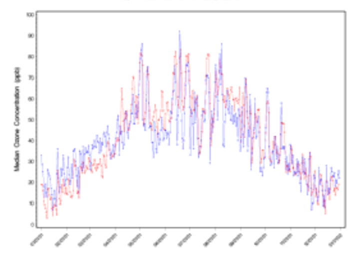
Objectives

- Provide daily particulate matter ($PM_{2.5}$) and ozone (O_3) spatial surfaces for Environmental Health Tracking
- Combined predictions can be used for modeling air quality – public health relationships in the Public Health Air Surveillance Evaluation (PHASE) project
- Determine air quality non-attainment areas

Data Sources

- 24-hr average $PM_{2.5}$ data from EPA's FRM fine Particulate Network
- Daily 8-hr maximum O_3 concentrations from the NAMS/SLAMS Network
- Community Multi-Scale Air Quality (CMAQ) daily $PM_{2.5}$ and 8-hr maximum O_3 output over 36 km grid
- MODIS Satellite Aerosol Optical Depth (AOD) data over 10 km grid
- Eta Data Assimilation System (EDAS) meteorological data over 80 km grid
- LandScan daytime population density data
- 24-hr average $PM_{2.5}$ data from EPA's Speciation Trends Network (STN) and Interagency Monitoring of Protected Visual Environments (IMPROVE) data for validation
- Daily 8-hr maximum O_3 concentrations from EPA's Clean Air Status and Trends Network (CASTNet) for validation





Methods

- Monitoring data, CMAQ output, and MODIS data can be used simultaneously to predict daily pollutant surfaces
  - Air quality monitoring data is spatially sparse, temporally rich
  - Numerical model output has high spatial and temporal resolution, but potential for location dependent bias
  - Satellite data has high spatial and temporal resolution, but potential for location specific bias and significant missing data (cloud cover)
- Leads to more accurate predictions and prediction errors
- Draw on strengths of each data source:
  - Give more weight to accurate monitoring data in areas where monitoring data exists
  - Rely on bias adjusted model output and satellite data in non-monitored areas
- Model underlying spatial dependence and measurement errors of each data source – no blind combining
  - Monitor data

$$\boldsymbol{X}_t^k(s_{ij})\,|\,\boldsymbol{W}_t(s_{ij}),\sigma_X^2 \sim N(\boldsymbol{W}_t(s_{ij}),\sigma_X^2)$$

  - CMAQ Data

$$\boldsymbol{Y}_t^k(s_{ij})\,|\,\boldsymbol{W}_t(s_{ij}),\boldsymbol{\beta}_D,\sigma_X^2 \sim N(\boldsymbol{W}_t(s_{ij})+\boldsymbol{D}_t(s_{ij})\boldsymbol{\beta}_D,\phi\sigma_X^2)$$

  - AOD Data

$$\boldsymbol{S}_t^k(s_{ij})\,|\,\eta,\boldsymbol{V}_t(s_{ij}),\sigma_S^2 \sim N(\eta+\boldsymbol{V}_t(s_{ij}),\sigma_S^2)$$

  - Underlying air quality process

$$\boldsymbol{W}_t(s_{ij})=\mu+\boldsymbol{A}_t(s_{ij})\boldsymbol{\beta}_A+\boldsymbol{Z}_t(s_{ij})$$

  - Air quality process residuals and underlying AOD process includes an auto-regressive temporal component and a conditionally auto-regressive spatial component

$$\boldsymbol{V}\,|\,\sigma_V^2,\rho_V \sim N\left(0,\sigma_V^2(\boldsymbol{\Lambda}_\tau^{-1}(\rho_V)\otimes\boldsymbol{\Lambda}_p^{-1})^{-1}\right) \qquad \boldsymbol{Z}\,|\,\sigma_Z^2,\rho_Z \sim N\left(0,\sigma_Z^2(\boldsymbol{\Lambda}_\tau^{-1}(\rho_Z)\otimes\boldsymbol{\Lambda}_p^{-1})^{-1}\right)$$

- Hierarchical Bayesian statistical modeling based on custom-designed Monte Carlo Markov Chain software

Results and Future Work

Results

- Combined approach provides reliable information about the true $PM_{2.5}$ and O_3 surfaces.
- How well does the combined approach predict to NON-MONITORED locations? Answer: Validate the model against data not used in fitting the model, use IMPROVE and STN $PM_{2.5}$ and CASTNet O_3 data.
  - Calculate root mean squared prediction error (RMSPE)
  - Compare predictive results of combined approach to ordinary kriging

Future Work

- Use 12 km CMAQ gridded output and compare to current results with 36 km output
- Use model to refine definition of pollution non-attainment areas

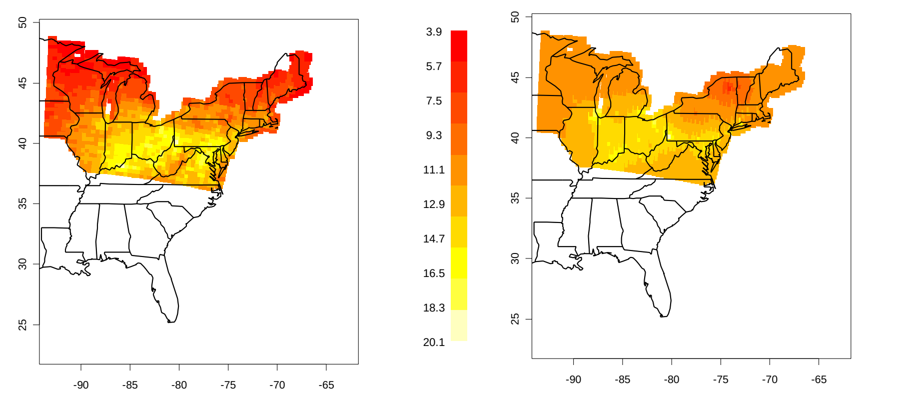


Figure 1. Predicted summer average $PM_{2.5}$ (*$\mu g/m^3$*) surface – Combined model (left) versus interpolated monitoring data (right)

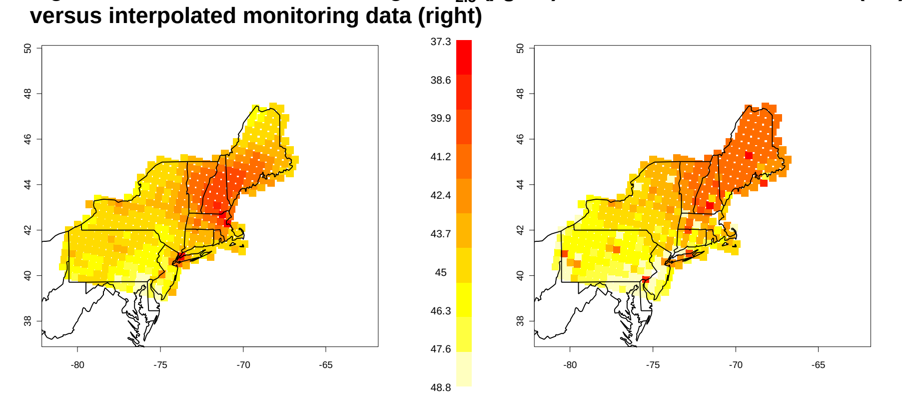


Figure 2. Predicted O_3 (*ppb*) seasonal average surface – Combined model (left) versus interpolated monitoring data (right)

Table 1. Root Mean Squared Prediction Errors (RMSPE) Using Three Prediction Surfaces

	O_3			$PM_{2.5}$		
		Combined Model Improvement (%)			Combined Model Improvement (%)	
Prediction Surface	Overall RMSPE	Sites 11	Days 245	Overall Estimate	Sites 60	Days 78
CMAQ Output	0.525	91%	64%	0.277	90%	87%
Kriged Monitor Data	0.521	91%	57%	0.165	58%	69%
Bayesian Combined	0.501			0.127		

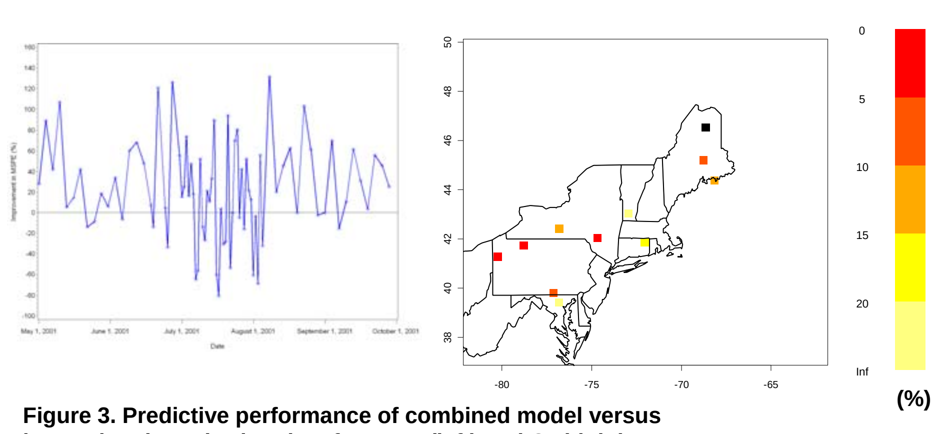


Figure 3. Predictive performance of combined model versus interpolated monitoring data for $PM_{2.5}$ (left) and O_3 (right)